

# **National Digital Library Program Text Keying And Encoding Instructions**

**Version 97-1**

March 12, 1997

The Library of Congress  
Washington, DC

This document was compiled and edited by the National Digital Library Program, Library of Congress. It is intended for documents that are encoded with the American Memory DTD revised in February 1997(ammem2.dtd). Please refer questions and suggestions to the Text Quality Review Committee: Martha Anderson, Tom Bramel, Beth Davis-Brown, Judith Davis, LeeEllen Friedland, and Juretta Hecksher.

**Revisions:**

The following sections were revised March, 1997:

I.3.2.

Example B.1.

Example D.2.

The following sections were revised September, 1997.

B.3.1.

H.1.

I.4.3

Clarifications 4/97

Section AKeying .....	1
A.1. General Instructions.....	1
A.2. Text And Marks To Key.....	1
A.3. Text And Marks That Will Not Be Keyed Or Retained .....	1
A.4. Special Characters And Layout.....	2
Section BNaming .....	3
B.1. Targets .....	3
B.2. Document Naming.....	3
B.3. Naming Of References .....	3
Section CTagging.....	5
C.1. Insertion Of Tags.....	5
C.2. Spacing .....	5
C.3. Page Breaks.....	5
C.4. Page Numbers .....	5
C.5. Blank Pages .....	6
C.6. Line Breaks .....	6
C.7. Catch Words .....	6
Section DStructural Elements .....	7
D.1. Document Components .....	7
D.2. Header.....	7
D.3. Text.....	7
D.4. Front Matter.....	7
D.5. Main Body.....	7
D.6. Back Matter .....	8
D.7. Headings .....	8
D.8. Divisions .....	8
D.9. Division Type Attributes.....	8
D.10. Paragraphs.....	9
D.11. Emphasis .....	9
D.12. Block Indents .....	9
D.13. Superscript And Subscript .....	10
Section ESpecial Text .....	11
E.1. Advertisements.....	11
E.2. Deleted Text.....	11
E.3. Handwritten Text .....	11
E.4. Added Text .....	11
E.5. Unkeyable Text .....	12
E.6. Stamped .....	12

E.7.    Fractions .....	12
Section FSpecial Document Instructions.....	13
F.1.    Document Sets .....	13
F.2.    Specified Elements.....	13
F.3.    Specified Attributes .....	13
Section GNotes And Anchors.....	14
G.1.    Notes.....	14
G.2.    Anchors.....	14
G.3.    Anchor Attributes.....	14
G.4.    Note Attributes.....	14
Section HIllustrations.....	15
H.1.    Illustrations.....	15
Section ITables And Lists.....	16
I.1.    Tables.....	16
I.3.    Lists .....	16
I.4.    Type Attribute For Lists .....	16
I.5.    Simple Lists.....	16
I.6.    Lists Vs. Tables .....	17
Section JSpecific Page Types .....	18
J.1.    Title Pages.....	18
J.2.    Letterhead .....	18
J.3.    Bookplates .....	18
J.4.    Targets .....	18
J.5.    Forms .....	18
J.6.    Table Of Contents .....	19
Section KQuality Review And Delivery .....	20
K.1.    Vendor Quality Review .....	20
K.2.    Delivery Of Completed Document Texts.....	20
Appendix of Examples.....	21



## **SECTION A KEYING**

### **A.1. GENERAL INSTRUCTIONS**

1. Unless otherwise instructed, key all words in the document, left to right, top to bottom. Words are to be keyed in intelligent clusters. For example, text in each cell of a table should be keyed as a unit, rather than reading across a row and concatenating words in different table cells. See Example A.1.1. See also Section I, Tables and Lists.
2. Words are to be keyed exactly as they appear. Retain all the variant and incorrect spelling in the original text. For exceptions to this rule, see Section A.3, Text And Marks That Will Not Be Keyed Or Retained.
3. Columnar text will be treated as flowing text. Key first column followed by second column, etc. Refer also to Section I.1 for keying tables. See Example A.1.3.
4. Footnotes will be keyed at the end of the paragraph of the first reference. Endnotes will be keyed where they appear. Key margin notes immediately following its closest paragraph. See section G.1 for tagging of notes.

### **A.2. TEXT AND MARKS TO KEY**

Key the following text features:

1. Only the first occurrence of letterhead
2. Text of advertisements, unless Document Instructions say to omit advertising text. For complicated advertising formats, key the text as table text in cells.
3. Masthead of a newspaper, telegram, etc.
4. Stamped, embossed, and perforated marks
5. Page numbers
6. Captions of illustrations
7. Text of bookplates

### **A.3. TEXT AND MARKS THAT WILL NOT BE KEYED OR RETAINED**

Do not key:

1. Running heads
2. Text in illustrations
3. Telephone book-style "ears"
4. Hyphens that appear only because a word was too big to fit on a line (Note: When a word is hyphenated as the last word on the page, complete the word before beginning the page information group tags.)
5. Letterhead (heads of forms or personal printed stationery) except for each first appearance
6. Immediate corrections. (Note: Where typos have been struck over, key the corrected letter and ignore the wrong letter.)

7. Incidental marks such as coffee stains, blood, doodles, fingerprints, etc.
8. Rules, vines, borders, and other decorations
9. Text bleedthrough from reverse of page.

#### A.4. SPECIAL CHARACTERS AND LAYOUT

1. For non-ASCII characters, e.g. § (section symbol), ° (degree symbol), & (ampersand), † (dagger) etc., key the appropriate character entity. For example, **&sect;**, **&degree;**, **&amp;**, **&dag;**. Refer to ISO 8879 for publicly declared character entities. If there is no publicly declared entity, key three question marks inside square brackets. For example, [???]
2. Key line breaks wherever the text is ended before the customary margin for a document as on the title page of a book or in poetry.
3. Replace leader dots and other graphic connectors with an <hsep> tag. See Example A.4.3.
4. Key ellipses as a series of periods.
5. When braces group items, key all items on the left of the braces, then key items on the right. If there are one or two groupings, tag as a list. If three or more groupings, tag as a table. Do not key the brace character. See section I. 6. See example A.4.5.
6. Illuminated characters and other odd-sized or decorated letters should be tagged as <hi rend="other">. The entire word should appear between the tags, not just the initial letter. See Example A.4.6. *Encoding example:* <hi rend="other">That</hi>
7. When a word has more than one form of highlighting or emphasis, such as italic and bold, the attribute value for the <hi> tag should be "other". The entire word should appear between the <hi> tags.

#### A.5. TYPOGRAPHICAL DESIGN OF ORIGINAL

1. Do not try to mimic the typographical design or format of the original by using extra hard returns, spaces or other typing conventions.
2. Do not try to capture decorative fonts and styles on title pages or in headings. See example A.5.2.
3. Special Document Instructions will be provided with document sets that may contain non-twentieth century printing conventions and text that is oriented in various directions on the same page.

## SECTION B NAMING

### B.1. TARGETS

Each document to be scanned is preceded by an identification target. The target should be the first scanned image for each set of document images. Identification targets are always numbered 0 with as many leading zeroes as required to create the minimum digits for the filename. The target has all the information necessary to create the <teiheader>.

There may also be additional scanning information provided below the horizontal line on the target. Do not key the line or anything below the line.

The <amid> element of the <teiheader> contains the item identifier for the document. In the following example, <amid type="aggitemid">rbnawsa-n8358</amid>, the item identifier is n8358. See Example B.1.

### B.2. DOCUMENT NAMING

1. The filename for the SGML-encoded, machine-readable text will be the item identifier followed by the extension sgm. It is stored in a directory named for the item identifier.

<b>Directory name</b> from identification target item identifier.	n8358
Converted/marked-up <b>document</b> filename	n8358\n8358.sgm
Identification <b>target</b> image filename	n8358\0000.tif
<b>1st page image</b> filename	n8358\0001.tif
<b>17th page image</b> filename	n8358\0017.tif

### B.3. NAMING OF REFERENCES

1. References to external files are designated with the ENTITY attribute of the element. ENTITY references are used with <controlpgno>, <illus>, and <table> elements. For <controlpgno> and <table>, the ENTITY value consists of the page image filename without the extension preceded by the letter p. For the ENTITY value of <illus>, the filename without extension is preceded by the letter i.
  - a. The contents of the identification target image (0000.tif) are used in the <teiheader> only. The image is not referenced in the text.
  - b. 1st page image is named 0001.tif. The <controlpgno> ENTITY value is p0001. Type the actual number, 0001, between the start and end <controlpgno> tags. *Encoding example:*  
<controlpgno entity="p0001">0001</controlpgno>
  - c. 17th page image is named 0017.tif. The <controlpgno> ENTITY value is p0017. Type the actual number, 0017, between the start and end <controlpgno> tags. *Encoding example:*  
<controlpgno entity="p0017">0017</controlpgno>



- d. For an illustration appearing on control page 0017, the <illus> ENTITY value is 0017 preceded by the letter I. *Encoding example:*  
<illus entity="i0017">
  - e. For a table appearing on control page 0003, the ENTITY value of the <table> element is p0003. *Encoding example:*  
<table entity="p0003">
2. External references that point to files which are supplementary to the document are tagged with <xref>. The DOC attribute value is the entity reference to the external file. The use of this tag and the scheme for assigning the DOC value will be designated in the Document Instructions.
3. Internal references that do not refer to external files are designated with an ID attribute. The <anchor> of a note uses the ID attribute. The corresponding target in the <note> uses the ANCHOR.IDS attribute.
- a. To name the ID for the <anchor> element, always start with n (for note), followed by the control page number (padded with zeroes to make a four digit number), followed by a hyphen, followed by 01, if it's the first or only note on that page. *Encoding example:* <anchor id="n0019-01">  
If it is the second note on that page, it will be n0019-02. Type the actual reference character or entity (e.g., \*, 1, or &dag;) in between the start and end <anchor> tags.
  - b. For the corresponding <note> element, the ANCHOR.IDS value should match exactly the ID value in the anchor tag. *Encoding example:* <note anchor.ids="n0019-01"> Subsequent ANCHOR.IDS for an established note should be numbered sequentially in the regular manner. Type the actual reference character (e.g., \*, 1, or &dag;) -- if it appears before the note text, at the beginning of the note text after the start tag.

## SECTION C

### TAGGING OF PHYSICAL FEATURES

#### C.1. INSERTION OF TAGS

1. Tags must never be inserted into the middle of a word.
2. Tags must never replace a space between words.
3. All element names must be lower case.

#### C.2. SPACING

1. Gaps in text, where items are not tabular but are deliberately and clearly separated by various amounts of white space, should be marked by the `<hsep>` tag. The `<hsep>` tag is used to show significant amount of horizontal space between two portions of text. A blank line used to indicate space where names should be filled in (as on a form, for example), should be tagged as `<hsep>`. Horizontal lines that are simply a design should not be tagged as an `<hsep>`. See Example C.2.1.
2. Spaces in between the letters of a word should not be encoded. The text should be tagged as `<hi>` with the `REND` attribute value of "other" except when appearing in a title or heading. *Encoding example:* `<hi rend="other">CONGRESS</hi>` See Example C.2.2.

#### C.3. PAGE BREAKS

Every page break is marked with a set of `<pageinfo></pageinfo>` tags.

The `<pageinfo>` element contains `<controlpgno>` and `<printpgno>` elements.

`<controlpgno>` element captures the sequence number of the page within its document set and the `<printpgno>` captures the actual page number that appears on the page. See Section C.4., Page Numbers.

#### C.4. PAGE NUMBERS

1. Sequence of pages
  - a. The sequential number of the page images in the document (excluding blank pages), starting from 1, will be recorded in the `<controlpgno>` element.
  - b. The `<controlpgno>` element must have an `ENTITY` attribute set to `cccc` where `cccc` is the filename of the document. Control page numbers start at 1 for each document set, are front-filled with zeroes to the appropriate number of digits, and increment by 1. Control page numbers are independent of the print page number. The text within the `<controlpgno>` tag should be `cccc`. *Encoding example:* `<controlpgno ENTITY="0001">0001</controlpgno>`
  - c. `<controlpgno>` should be keyed at the beginning of a page, but not mid-word. If a word is split by a hyphen and the second part of the word

appears on the next page, the <controlpgno> tag should be inserted after that word.

2. Print page numbers
  - a. The actual page number printed on the page will be tagged with the <printpgno> tag within the <pageinfo> element.
  - b. When tagging a page number, keep the number and discard any characters such as brackets, braces, or the word "page" that are used to set off the number. For example, all the following would be tagged as <printpgno>3</printpgno>:  
PAGE:3      -3-      {3}      [3]      -page 3-
  - c. If there is more than one page number appearing on the page, key all page numbers using as many <printpgno> tags as necessary.
  - d. An unnumbered page is indicated by empty <printpgno></printpgno> tags, with no space between the start and end tags.

#### **C.5. BLANK PAGES**

1. Tagged as a regular page with a <blankpage> tag keyed into the <pageinfo> tag. The pageinfo that contain the <blankpage> tag will be followed immediately by the next <pageinfo> tag. *Encoding example:*  
<pageinfo><controlpgno entity="0000">0000  
</controlpgno><printpgno></printpgno><blankpage>  
</pageinfo>
2. The requirement for use of <blankpage> tags in a document set will be indicated in the Document Instructions. Only key the <blankpage> tag for the indicated pages.

#### **C.6. LINE BREAKS**

Structures that have embedded hard returns should have a line break tag (<lb>) keyed for the hard return. Embedded hard returns are implied when the line ends before the customary right margin of the document. The <lb> tag will most often be used to indicate hard returns on the title page or for significant structures such as poetry. See Example C.6.

#### **C.7. CATCH WORDS**

The odd words repeated at the end of a column or page of text to indicate the first word on the next column or page, will be treated as a new line of text, preceded and followed by the line break tag <lb>.

#### **C.8. TITLE PAGES**

Key line breaks on title pages marking them with <lb>. Do not tag emphasis or special fonts on title pages.

## SECTION D STRUCTURAL ELEMENTS

### D.1. DOCUMENT COMPONENTS

Documents conforming to the American Memory DTD (ammem.dtd) have two main components: <teiheader> and <text>.

### D.2. HEADER

1. The first scanned image for every document should be the target. The target is always numbered 0 (with as many leading zeroes required to create a minimum four-digit filename). The contents of the target should be used to create the <teiheader> that appears at the beginning of each converted document. See Example D.2. The target may contain additional scanning information below a horizontal line. Do not key the horizontal line or any information below the line.
2. Header attributes:
  - a. The "creator" attribute for the <teiheader> should read: "Library of Congress"
  - b. The "date.created" attribute for the <teiheader> should be set to current date.

See sample target, Example D.2.

### D.3. TEXT

1. The <text> element immediately follows the <teiheader> and contains the tagged document.
2. The National Digital Library Program uses only two text TYPE designations: publication or manuscript. The Library will specify which text TYPE is appropriate for each collection or set of documents. This information is generally provided on document targets following header contents.

### D.4. FRONT MATTER

Data before the main content of a document should be tagged with <front>. Front matter is indicated by the presence of headings such as table of contents, introduction, preface, dedications, foreword, bibliography, index, references, appendices, glossary, and publisher's notes. Actual text of headings may vary slightly. Contents of front matter may appear similar to back matter, i.e. an index may precede the main content of the document. *Encoding example:* <front><div><head>PREFACE.</head>...

### D.5. MAIN BODY

The main contents of a document should be tagged with the <body> element. The body of the document starts with regular pagination (if Front Matter has different pagination), contains regular paragraphs, and/or has text set off from front matter by horizontal

separator.

See Example D.5. *Encoding example:* <body><div><head>First Chapter<lb>WOMAN'S POSITION IN THE PAST.</head>...

## **D.6. BACK MATTER**

Data occurring after the main contents of a document should be tagged with the <back> element. Back matter is indicated by the presence of headings, such as dedications, bibliography, index, references, appendices, glossary, publisher's notes, and conclusions. Actual text of headings may vary slightly. Contents of back matter may appear similar to front matter, i.e. a table of contents may follow the main content of the document.

*Encoding example:* <back><div type="bib"><head>Bibliography</head>...

## **D.7. HEADINGS**

Headings should be tagged with the <head> element. Headings such as chapter or section heads are indicated by off-set text with uniform emphasis. Headings often appear in a larger type-face and are uniformly bold or italic or of a different font than the rest of the text. Do not tag the uniform emphasis. Tag any emphasis which is not part of the uniform emphasis such as a single italic word within the head. See section D.12, Emphasis.

## **D.8. DIVISIONS**

All documents must contain at least one division. Headings usually indicate divisions. Every heading must be preceded by a <div> tag. (Note: Some documents may have divisions that are not readily recognized by headings. When this is the case, rules for division recognition will be indicated in the Document Instructions.)

## **D.9. DIVISION TYPE ATTRIBUTES**

1. The <div> that is the most complete description of the document (title, author, copyright information, etc.) should have a TYPE attribute value of "idinfo". This type of division most commonly appears only once within a book, and usually as the title page within the front matter of a document. No headings should be tagged within this type of division and it may not contain any other division within it. (It does not nest.) The division could appear within the main body of a text, and even in the back matter. If more than one idinfo division is present in a volume, it will be indicated by the Library on the target. See Example D.9.1.
2. If the text in the division headings is one of the following, then the TYPE attribute should contain the parenthesis value. bibliography (bib); glossary (gloss) index (index); list of illustrations (listill); end notes (end notes); and table of contents (toc); (Actual text of headings may vary slightly.) If none of the headings are used, leave out the TYPE attribute.
3. The <div> element may carry an ID attribute. The Document Instructions which accompany a document set will indicate the requirement for this attribute, as well as the scheme for assigning IDs.

## D.10. PARAGRAPHS

1. Tag normal paragraph-sized units of text with the <p> element. A paragraph may be made up of incomplete sentences, and it may or may not be indented. It will appear uniformly within a document and should be tagged as such. Do not capture changes in font or line spacing.
2. Paragraphs ending with a colon or colon/m-dash. Use care in placement of the end </p> tag for paragraphs ending with either a colon (:) or a colon and an m-dash (:—) followed by a list. Tag the list and end the paragraph after the close list tag. Note: If these paragraphs are not followed by a list, end the paragraph normally,
3. Paragraphs that contain line breaks. If hard returns appear within a paragraph, the paragraph structure should be kept open until the end of the entire structure. A line break tag <lb>, should be used to indicate the hard return.
4. When indentation is unclear end the paragraph and begin a new paragraph when end punctuation, such as period, exclamation point, or question mark; is followed by a hard return and an indent. See example D.10.4.
5. The <p> tag is also used to encode text contained within the <item> and <caption> elements.  
*Encoding example:*<caption><p>Distant view of Mount Rushmore</p><caption>

## D.11. EMPHASIS

1. Emphasized text can usually be recognized by its different appearance from surrounding text. Text should be tagged for emphasis using the REND attribute on the <hi> element.
2. Specific types of emphasis to be identified (with the REND value indicated within the parentheses) are **bold** (bold), *italics* (italics), underline or double underline (underscore), handwritten underline (underscore), and SMALL CAPS (smallcaps). All other types of emphasis should be indicated with the REND value of "other".
3. If only a portion of a word is emphasized, the entire word should be tagged with the <hi rend="other"> element. For example, in some documents the first letter of each chapter is larger and more ornate than the rest of the word.
4. If more than one type of emphasis is used in a word, the entire word should be tagged <hi rend="other">. See Example D.11.4.
5. If spaces occur within a word, the spaces should not be captured, but the text should be tagged within <hi rend="other">.
6. Do not use <hi> within tables or headings or on title pages. See Section A.4.7.

## D.12. BLOCK INDENTS

Indented text should be tagged as <hi> like other emphasized text. The REND attribute

value is “blockindent”. When indented text occurs within a paragraph, place the end `</p>` tag at the end of the paragraph not before the indented text. See Example D.12.

#### **D.13. SUPERScript AND SUBScript**

Text appearing above the line should be tagged with `<superscript>`. Text appearing below the line should be tagged with `<subscript>`. If only a portion of a word is superscript or subscript, the entire word should be tagged with the appropriate element.

## SECTION E

### SPECIAL TEXT

Note: The following tags are subjective. The Library will therefore review these carefully. When in doubt, tag.

#### E.1. ADVERTISEMENTS

Tag advertisements with <ad>. Advertisements can often be recognized as portions of the text that are clearly an interruption of the normal text flow. Examples include an announcement of an event, a listing of products available, or a listing of services available. Advertisements may appear anywhere within a document. Advertisements often contain illustrations and may be separated from normal text flow by lines or boxes.

#### E.2. DELETED TEXT

Text that has been marked for deletion in the document should be tagged with <del>, using the REND attribute to indicate how the deletion is shown. Values for the REND attribute are “overstrike,” “erasure,” or “cancelled.”

#### E.3. HANDWRITTEN TEXT

1. It is important to capture the occurrences of handwritten material whenever they appear, regardless of their legibility. Handwritten text will be captured as follows:
  - a. Tag legible text within <handwritten> tags.
  - b. Tag illegible text as omitted within <handwritten> tags. See section E.5., Unkeyable Text.
  - c. Tag handwritten underlined text as <hi rend=“underscore”> See section D.11., Emphasis.
2. When the entire text of a document is handwritten, use <text type= “manuscript” rend= “handwritten”>. This information will be provided on the identification target images for each document following the <teiheader> element. Exceptions to this rule will be noted in the Document Instructions which accompany a document set.

#### E.4. ADDED TEXT

1. Any text that appears on the page that is not part of the flowing text and has an insertion point or some other indication of where it should appear will be tagged as <added>. The text itself should be keyed after the paragraph nearest the text. The PLACE attribute should be used to indicate where the added text appears on the page. Values for the PLACE attribute are “top”, “bottom”, “margin”, or “interlinear”.
2. Any added text for which an insertion point is not indicated should be keyed as notes. See section G, Notes and Anchors.



### **E.5. UNKEYABLE TEXT**

1. Any text which cannot be keyed should have an <omit> tag keyed with the attributes reason and extent to indicate why it could not be keyed and approximately how much data could not be keyed. The REASON attribute should be used to indicate why the text is omitted. Values for REASON are “illegible,” “missing,” or “untranscribable.” See Example E.5.1.
2. If the unkeyable text is less than one word, a question mark should be used to replace each unkeyable character. *Encoding example:* ba??n

### **E.6. STAMPED**

Any text which is part of text that has been stamped onto the hard copy should be tagged within <stamped> tags. Perforated or embossed text may also be tagged as <stamped>.

### **E.7. FRACTIONS**

When an ISOnum entity exists, use it to capture fractions. Example:  $\frac{1}{2} = \&frac12$ ; or  $\frac{1}{4} = \&frac14$ ; . If no publicly declared entity exists, key the fraction in the following manner.  $\frac{33}{100} = 33\&sol;100$ . If the fraction follows a whole number, key a space between the whole number and the fraction string. Example:  $4 \frac{33}{100} = 4\ 33\&sol;10$

## **SECTION F**

### **SPECIAL DOCUMENT INSTRUCTIONS**

#### **F.1. DOCUMENT SETS**

There are series of documents that require instructions from the Library regarding use of the specified elements or attributes. A description of the document set and special instructions will accompany the first shipment of the documents.

#### **F.2. SPECIFIED ELEMENTS**

These elements will be used only when specified by the Library for a defined document set.

1. Dates will NOT be tagged unless indicated in the Document Instructions. The Document Instructions will indicate how to identify and tag date using the `< date>` element.
2. External references when used will be specified in Document Instructions or other materials furnished with the document set. `< xref>` and `< xptr>` elements will be used for these references. The values for attributes and the position of these elements will be fully described in the Document Instructions.

#### **F.3. SPECIFIED ATTRIBUTES**

Full instructions for these attributes will be defined by the Document Instructions.

1. The use of ID attributes on some elements and the value scheme for assigning the ID.
2. Multiple occurrences of the requirement to use the IDINFO attribute on `< div>` .

## SECTION G

### NOTES AND ANCHORS

#### G.1. NOTES

1. Footnote text, referenced in the document text and printed at the bottom of the page, will be tagged as `<note>` and incorporated into the document text after the paragraph in which it is referenced. The `<anchor>` tag will be used to mark the reference to the footnote where it occurs in the document text.
2. Endnote text, referenced in the document text but printed at the end of a major division such as a chapter, will be tagged as `<note>` and incorporated into the document text at the division end. The `<anchor>` tag will be used to mark the reference to the endnote where it occurs in the document text.
3. Margin text, referenced in the document text with no indication of an insertion point, will be tagged as `<note>`. Key the margin note immediately following its closest paragraph.  
See example G.1.

#### G.2. ANCHORS

An anchor is a reference to any footnote, endnote, or margin note (that does not have an indication for its insertion point) indicated anywhere on the page. The `<anchor>` (reference location) gets an ID attribute. The `<note>` will be tagged with an ANCHOR.IDS attribute. Any margin note that has an indication to its insertion point will be tagged as `<added>`.  
See section E.4., Added Text .

#### G.3. ANCHOR ATTRIBUTES

The ID attribute of the `<anchor>` tag will be the `ncccc-##` where `cccc` is the controlpgno (front-filled with zeroes to 4 digits) and `##` is a sequential number front-filled with zeroes to 2 digits, starting at 01 on each page. Note: Multiple references to the same note will have different IDs.

*Encoding example:* `<anchor id="n0001-01">1</anchor>` represents the first reference to the first note on page 1 of the document.

#### G.4. NOTE ATTRIBUTES

1. The ANCHOR.IDS attribute of the `<note>` tag will be a listing of all the ANCHOR.IDS that represent that note. Each ANCHOR.IDS value will be followed by a space (except the last one). *Encoding Example:* `<note anchor.ids="n0001-01" anchor.ids="n0030-02">`
2. The location of the text of the note should be indicated within the PLACE attribute. Values for the PLACE attribute are "top," "bottom," "margin," or "interlinear."

## SECTION H ILLUSTRATIONS

### H.1. ILLUSTRATIONS

Non-textual material with a corresponding page image file should be tagged as an illustration with the <illus> tag. The associated caption should be keyed within the <caption> tag. An ENTITY attribute will be used to indicate the pointer to the corresponding image file. The attribute will be the filename without the extension preceded by a feature designator i.

*Encoding example:* An illustration appearing in the image file 0017.tif, should be tagged <illus entity="0017i"><caption><p>Illustration X.</p></caption></illus> See example H.1.

## SECTION I TABLES AND LISTS

### I.1. TABLES

1. The purpose of capturing tables is for text searching only . The only information to be captured for tables is the title and each cell, in sequence, from left to right, top to bottom. Tag the title of a table as a caption. See example I.1.
2. Typographical composition of the tables should not be captured. Do not adjust for spanning or alignment. Do not key empty cells. Do not key any emphasis.
3. An ENTITY attribute will be used as a pointer to the page image for the table. The attribute value will be the page image filename without the extension.  
*Encoding Example:* <table entity="0017"><caption><p>Table of States</p></caption><cell>State</cell><cell>Capital</cell><cell>Flower</cell><cell>South Carolina</cell><cell>Columbia</cell><cell>Jasmine</cell></table>

### I.2. CAPTIONS INSIDE TABLES

A heading that is positioned over a table or near an illustration should be tagged with the <caption> element.

### I.3. LISTS

1. Any itemization is tagged as a list. This includes numbered paragraphs, bulleted paragraphs, tables of contents, indexes, paragraphs with hanging indents, etc.
2. If a list of numbers is followed by a total line, the last number in the column above the line should be tagged with <hi rend="underscore">.
3. If a list is bulleted, capture the bullet regardless of appearance with the &bull; entity.

### I.4. TYPE ATTRIBUTE FOR LISTS

Lists are of three types:

1. Sequenced with numbers, letters, roman numerals, etc. (TYPE ="ordered")
2. Bulleted with stars, dashes, circles, bullets, pointing hands, etc. (TYPE ="bulleted") Key bullet with character entity &bull;
3. Simple(See section I.5., Simple lists.)

### I.5. SIMPLE LISTS

Lists that are not sequenced or bulleted can be identified in a number of ways:

1. Hanging indents
2. Homogeneous information sometimes listed in 2 or more columns
3. Table of Contents
4. 2 columns of information without a heading that describes each column

### I.6. LISTS VS. TABLES

1. Tables are defined as 3 or more columns of information with headings of some kind at the top of each column. 1 or 2 column tables are to be keyed as a list with the <hsep> to set the data apart. A graphic separator of data (like line drawing) would indicate that the structure is a table. Table of Contents and Indices are always lists.
2. Braces grouping items together will be keyed as tables except in cases where curly braces are used in all or part of a two-column list. See Document Instructions for keying braces as part of a list. See example I.6.2.

## SECTION J

### SPECIFIC PAGE TYPES

#### J.1. TITLE PAGES

Key the text on title pages using paragraph tags to indicate logical groupings of information. For example, for a centered title and author statement on a title page, begin with `< p>` , type the text using `< lb>` to indicate where the lines end, and close the paragraph `< /p>` when the statement is complete. Using this approach, most title pages are likely to have at least one paragraph containing the title and author information and another paragraph containing the publication information.

#### J.2. LETTERHEAD

Every time there is letterhead that is not identical to that on the previous page (of a letter, for example), it should be keyed and tagged as text.

#### J.3. BOOKPLATES

Key all the text contained in bookplates. Use the linebreak element `< lb>` to separate short lines of text.

#### J.4. TARGETS

1. Do not treat targets as the first page of a document. (Page images of targets should always have filenames that end with at least two zeroes, "00".)
3. The text provided on the target should be keyed in the appropriate part of the document `< teiheader>` . Most targets will contain the text for the entire `teiheader`.

#### J.5. FORMS

1. A form is defined as preprinted questions or statements where a user response is required. A form generally contains at least one blank line or space that is used for filling in information.
2. The information supplied by the respondent does not stand alone; therefore both the full text of the "question" and the "answer" must be keyed and tagged;
3. The boxes and blank lines on the form should not be keyed;
4. Since images of the pages will always be supplied, it is not necessary to distinguish explicitly between the "question" and the "answer."

For example:

goat	( )
dog	(X)
cat	( )

could be keyed as:

```
< list> < item> < p> goat< /p> < /item>
```

```
< item> < p> dog< hsep> X< /p> < /item>
< item> < p> cat< /p> < /item> < /list>
```

Note: dog and X are separated by the < hsep> tag to indicate horizontal separation.

## **J.6. TABLE OF CONTENTS**

Table of contents pages should be keyed as lists. Insert < hsep> to replace leader dots between the title or description and the page number. See example J.6.

*Encoding example:*< list> < head> Contents< /head> < item> < p> I. OUGHT  
WOMEN TO LEARN THE ALPHABET?< hsep> 1< /p> < /item>

## **J.7. INDEXES**

Indexes should be keyed as lists. Items may contain paragraphs, illustrations, advertisements, lists, notes or tables. When these elements occur in the Index, they should be tagged appropriately.

See example J.7.



## **SECTION K**

### **QUALITY REVIEW AND DELIVERY**

#### **K.1. VENDOR QUALITY REVIEW**

1. Parse all files. All documents must conform to the American Memory DTD and be validated with three parsers.
2. Identify cropped page images that may result in incomplete keying. Flag instances of short or incomplete pages.
3. Check accompanying files for the sequence of page numbers; the correct format for entity references to page images, illustrations, and tables; and any occurrences of omitted text.

#### **K.2. DELIVERY OF COMPLETED DOCUMENT TEXTS**

Each document must be provided to the Library in a single file. If a document is broken into multiple parts for keying and/or tagging, it must be reassembled into a single file before delivery.

## **APPENDIX OF EXAMPLES**

## Appendix B

### NDLP Paper Scanning Contract

### Library of Congress

### Text Conversion Startup Review

---

April 14, 1997

#### Clarifications:

1. Use the doctype statement exactly as it appears on the target. `<!doctype tei2 public "-//Library of Congress - Historical Collections (American Memory)//DTD ammem.dtd//EN" [!entity...]>`. If necessary change the dtd filename to ammem.dtd. Our local configuration depends upon this doctype statement.
2. Insert the appropriate dates in the teiheader element. These are represented by YYYY/MM/DD. Y=year to 4 digits, M=month to 2 digits, D=day to 2 digits. Example: 1997/04/07.
3. Key targets exactly inserting date information . The LC will assume responsibility for any errors that are introduced by a faulty target.
4. The SGML declaration used with ammem.dtd does not allow SHORTTAG therefore attributes with default values must be fully expressed. Please key the following default attributes when appropriate:

For `<amcolid>` element, the TYPE attribute must be keyed with a default value of "aggid". This will be inserted in the teiheader information that appears on the target.

`<amcolid type="aggid">`

For `<illus>` elements, the MAP attribute must be keyed with a default value of "no". If the illustration is a map, then the value will be "yes".

`<illus entity="i0000" map="no">`

For `<list>` elements, the TYPE attribute must be keyed with a default value of "simple". `<list type="simple">` If the list is ordered or bulleted, use the appropriate attribute. (See Section I.4. of *Keying and Encoding Instructions* for description of list types.)

For `<omit>` elements, the REASON attribute must be keyed with a default value of "illegible".

`<omit reason="illegible" extent="6 words">`

If `<date>` elements are used, the CERTAINTY attribute must be keyed with a default value of "certain". Other values may be specified in special instructions accompanying

material to be keyed. <date value="yyyy/mm/dd" certainty="certain">

5. As DCL pointed out, entity values must begin with an alpha character. The following is a change to Section B.3. of the *Keying and Encoding Instructions*.

### B.3. NAMING OF REFERENCES

1. References to external files are designated with the ENTITY attribute of the element. ENTITY references are used with <controlpgno>, <illus>, and <table> elements. For <controlpgno> and <table>, the ENTITY value consists of the page image filename without the extension preceded by the letter *p*. For the ENTITY value of <illus>, the filename without extension is preceded by the letter *i*.
  - a. The contents of the identification target image (0000.tif) are used in the <teiheader> only. The image is not referenced in the text.
  - b. 1st page image is named 0001.tif. The <controlpgno> ENTITY value is p0001. Type the actual number, 0001, between the start and end <controlpgno> tags. *Encoding example:*  
<controlpgno entity="p0001">0001</controlpgno>
  - c. 17th page image is named 0017.tif. The <controlpgno> ENTITY value is p0017. Type the actual number, 0017, between the start and end <controlpgno> tags. *Encoding example:*  
<controlpgno entity="p0017">0017</controlpgno>
  - d. For an illustration appearing on control page 0017, the <illus> ENTITY value is 0017 preceded by the letter *I*. *Encoding example:*  
<illus entity="i0017">
  - e. For a table appearing on control page 0003, the ENTITY value of the <table> element is p0003. *Encoding example:*  
<table entity="p0003">
6. Please key catchwords, the words at the bottom of a page that indicate the first word on the following page. Examples are found in RB17.
7. Do not tag empty cells in table text. HJ01 page 754 (control page 0065) shows tagging of empty cells. See example Law A.
8. There is a clarification of how to key a type of two-column list for the Law text. Alphabetical lists of names (HJ01, page 157, control page 0026) appearing in two columns should be keyed as if they were newspaper columns. Key all of the left column, then all of the right column. See Law B for examples.